Journal of Multivariate Analysis **74**, 49–68 (2000)
doi:10.1006/jmva.1999.1874, available online at http://www.idealibrary.com on IDEAL®

Asymptotic Normality of Posterior Distributions for Exponential Families when the Number of Parameters Tends to Infinity

Subhashis Ghosal

Division of Mathematics and Computer Science, Free University, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Received November 12, 1997

We study consistency and asymptotic normality of posterior distributions of the natural parameter for an exponential family when the dimension of the parameter grows with the sample size. Under certain growth restrictions on the dimension, we show that the posterior distributions concentrate in neighbourhoods of the true parameter and can be approximated by an appropriate normal distribution.

AMS 1991 subject classifications: primary 62F15, 62J05.

Key words and phrases: exponential family, normal approximation, posterior consistency, posterior distribution.

1. INTRODUCTION

Exponential families arise naturally in statistical modelling and the maximum likelihood estimate (MLE) is consistent and asymptotically normal for these models [Berk [2]]. In practice, often one needs to consider models with a large number of parameters, particularly if the sample size is large; see Huber [14], Haberman [13] and Portnoy [18–21]. One may also think that the true model can only be approximated by a finite dimensional parametric model and the quality of the approximation improves with the dimension. In other words, we let the dimension of the parameter space grow with the sample size. Usual asymptotics of fixed dimension do not justify the large sample approximations in these situations and one needs more delicate results paying special attention to the increasing dimension. Consistency and asymptotic normality of the MLE in exponential families with an increasing number of parameters were established by Portnoy [21] under some conditions on the growth rate of the dimension of the parameter space. In this paper, we show that the

0047-259X/00 $35.00

posterior distribution can be approximated by a suitable normal distribution when the dimension increases to infinity. For fixed dimensional regular statistical models, the posterior distribution is asymptotically normal; see, for example, Le Cam [16], Bickel and Yahav [3], Johnson [15] and Ghosal *et al.* [10]. In Ghosal [7, 8], the present author showed that respectively for generalized linear models and linear regression models with number of regressors tending to infinity with the sample size, posterior asymptotic normality holds under a certain growth condition on the number of regressors. In models with an increasing number of parameters, justifying the asymptotic normality of the posterior distribution is more involved, since various constants appearing in the bounds for the error terms depend on the dimension. Thus some growth conditions on these constants are required, which in turn require some growth condition on the dimension. The exact requirement varies from example to example.

An important difference between our assumptions and those of Portnoy [21] is that the bound for the moments of the standardized observation are allowed to grow with the dimension. This introduces more flexibility and substantially broadens the applicability of the results. The bounds for the moments satisfy the required growth conditions if sufficiently strong growth condition on the dimension is imposed. However, a bound free of the dimension, as assumed in Portnoy [21], is usually not available. In the proof of the asymptotic normality of the posterior, we need to exploit consistency of the MLE. However, Theorem 2.1 of Portnoy [17] is inadequate for our purpose because it assumes a condition on the eigenvalues of the covariance matrix [Portnoy [21, Eq. (2.4)]] which is hard to check, in addition to assuming that the moments of the standardized variable are bounded by a constant independent of the dimension [Portnoy [21, Eq. (3.2)]]. In fact, Portnoy's [21] assumption (2.4) on eigenvalues fails to hold in the important example of multinomial distribution, and in general, whenever the minimum eigenvalue tends to zero. We therefore establish an alternative theorem on the consistency of the MLE avoiding assumptions (2.4) and (3.2) of Portnoy [21]. This result, stated as Theorem 2.1, is an important intermediate step for the approximation of the posterior and is believed to be useful to a frequentist also.

We organize the paper as follows. In Section 2, the setup is described and the main result on asymptotic normality of the posterior is proved. The aforesaid result on the consistency of the MLE is also presented in this section. Some auxiliary lemmas are used in the proof of the main theorem, whose proofs are given in the appendix. In Section 3, we apply our results to the multinomial distribution and a Bayesian density estimation problem. In Section 4, results of Section 2 are applied to the problem of estimation of the mean vector of an infinite dimensional normal distribution.

2. MAIN RESULTS

Suppose that for every n, we have a positive integer p_n, where $p_n \to \infty$ as $n \to \infty$, and p_n-dimensional independent random samples $\mathbf{x}_1^{(n)}, ..., \mathbf{x}_n^{(n)}$ from a p_n-dimensional standard exponential family with density

$$f(\mathbf{x}; \boldsymbol{\theta}_n) = \exp[\mathbf{x}^T \boldsymbol{\theta}_n - \psi_n(\boldsymbol{\theta}_n)], \tag{2.1}$$

where $\boldsymbol{\theta}_n \in \Theta_n$, an open subset of $\mathbb{R}^{p_n}$. We shall often suppress the index n to write p, $\boldsymbol{\theta}$, Θ, ψ and $\mathbf{x}_i$ for p_n, $\boldsymbol{\theta}_n$, Θ_n, ψ_n and $\mathbf{x}_i^{(n)}$ and respectively, but we keep in mind that all of these objects changing with n. Fix a parameter point $\boldsymbol{\theta}_0 \in \Theta$ which will be regarded as the "true parameter point". More precisely, since the true parameter changes with n, this is actually a sequence of parameter points. To prevent $\boldsymbol{\theta}_0$ approaching the boundary as $n \to \infty$, we assume that for a fixed $\varepsilon_0 > 0$ independent of n, the ball of radius ε_0 around $\boldsymbol{\theta}_0$ for the Euclidean distance is contained in Θ. All the probability statements, except when explicitly mentioned otherwise, refer to the parameter $\boldsymbol{\theta}_0$.

Set $\boldsymbol{\mu} = \psi'(\boldsymbol{\theta}_0)$ and $\mathbf{F} = \psi''(\boldsymbol{\theta}_0)$, the mean vector and the covariance matrix of the observations respectively. Note that $\mathbf{F}$ is also equal to the Fisher information matrix and is positive definite. Let $\mathbf{J}$ be a square root of $\mathbf{F}$, i.e., $\mathbf{J}\mathbf{J}^T = \mathbf{F}$. The MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is unique and satisfies $\psi'(\hat{\boldsymbol{\theta}}) = \bar{\mathbf{x}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i$.

Below, for a vector $\mathbf{x} = (x_1, ..., x_p)$, $\|\mathbf{x}\|$ will stand for its Euclidean norm $(\sum_{i=1}^p x_i^2)^{1/2}$. For a square matrix $\mathbf{A}$, $\|\mathbf{A}\|$ will stand for its operator norm defined by $\sup\{\|\mathbf{A}\mathbf{x}\|: \|\mathbf{x}\| \leqslant 1\}$.

Let, for $c \geqslant 0$,

$$B_{1n}(c) = \sup\left\{E_\theta |\mathbf{a}^T\mathbf{V}|^3 : \mathbf{a} \in \mathbb{R}^p,\ \|\mathbf{a}\| = 1,\ \|\mathbf{J}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 \leqslant \frac{cp}{n}\right\},$$

$$B_{2n}(c) = \sup\left\{E_\theta |\mathbf{a}^T\mathbf{V}|^4 : \mathbf{a} \in \mathbb{R}^p,\ \|\mathbf{a}\| = 1,\ \|\mathbf{J}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 \leqslant \frac{cp}{n}\right\},$$

where $\mathbf{V}$ is distributed as $\mathbf{J}^{-1}(\mathbf{U} - E_\theta \mathbf{U})$ and $\mathbf{U}$ has density (2.1). It may be noted that, since two square roots of a positive definite matrix are orthogonal multiples of each other and $\|\mathbf{J}\mathbf{u}\|^2 = \mathbf{u}^T\mathbf{F}\mathbf{u}$ for any vector $\mathbf{u}$, $B_{1n}(c)$ and $B_{2n}(c)$ are independent of the choice of the square root $\mathbf{J}$ of $\mathbf{F}$. To establish asymptotic properties, some growth conditions will be assumed on these numbers (see Condition (R) below). However, unlike Portnoy [17], we do not assume that the quantities $B_{1n}(c)$ and $B_{2n}(c)$ are bounded. As mentioned in the introduction, this relaxation is important since in examples, the bounds $B_{1n}(c)$ and $B_{2n}(c)$ also increase to infinity with the dimension.

It will be assumed that the prior satisfies the condition (P) below.

Condition (P). The prior distribution is proper, has a density $\pi(\cdot)$ which satisfies, at $\boldsymbol{\theta}_0$, the positivity requirement

$$-\log \pi(\boldsymbol{\theta}_0) = O(p \log p) \tag{2.2}$$

and Lipschitz continuity

$$|\log \pi(\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta}_0)| \leqslant K_n(c) \, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|$$
$$\text{for} \quad \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leqslant \sqrt{\|\mathbf{F}^{-1}\| \, cp(\log p)/n}, \tag{2.3}$$

where the Lipschitz constant $K_n(c)$ is subject to some growth restriction (see Condition (R) below).

Note that if the components of $\boldsymbol{\theta}$ are a priori independently distributed with the jth component θ_j following a density $\pi_j(\cdot)$, $j = 1, ..., p$, where for some M, δ, $\eta_0 > 0$ and for all $j = 1, ..., p$, $\pi_j(\theta_{0j}) > \eta_0$ and

$$|\log \pi_j(\theta_j) - \log \pi_j(\theta_{0j})| \leqslant M \, |\theta_j - \theta_{0j}|, \qquad |\theta_j - \theta_{0j}| \leqslant \delta, \tag{2.4}$$

then (2.2) and (2.3) are satisfied with $K_n(c) = Mp^{1/2}$ provided $\|\mathbf{F}^{-1}\| \, p(\log p)/n \to 0$.

The following condition on the growth rate of the aforesaid quantities will be assumed.

Condition (R). $B_{1n}(0) \, p^{3/2}(\log p)^{1/2}/\sqrt{n} \to 0$, $p \, \|\mathbf{F}^{-1}\|/n \to 0$; for all $c > 0$, $\sqrt{p/n} \, B_{1n}(c) \to 0$, $B_{2n}(c \log p) \, p^2(\log p)/n \to 0$ and $K_n(c) \sqrt{\|\mathbf{F}^{-1}\| \, p(\log p)/n} \to 0$; $\text{tr}(\mathbf{F})$ is bounded by a polynomial in p.

Since the determinant of a positive definite matrix is the product of its eigenvalues and the trace is the sum, it follows from the arithmetic mean-geometric mean inequality that

$$\det \mathbf{F} \leqslant (\text{tr}(\mathbf{F})/p)^p.$$

Therefore the growth rate of $\log \det \mathbf{F}$ is at most of the order $p \log p$.

In examples, quantities appearing in Condition (R), such as $B_{1n}(c)$, $B_{2n}(c \log p)$ and $\|\mathbf{F}^{-1}\|$ will grow like a power of p. Hence if n is sufficiently large compared to p, or equivalently, the growth of p with respect to n is sufficiently slow, then Condition (R) will hold. The exact requirement on the growth rate of p depends on the particular model under consideration. For the multinomial model, Condition (R) holds if $p^6(\log p)/n \to 0$ (see Section 3), whereas for the normal model, Condition (R) is satisfied if $p^3(\log p)/n \to 0$ (see Section 4).

In the proofs, we shall actually make the additional assumption that some power of p grows faster than n, and so $\log p$ and $\log n$ are of the same

order. When this condition fails but Condition (R) holds, we may split the integrals into regions $\|\mathbf{u}\| \leqslant n^{1/4}$ and $\|\mathbf{u}\| > n^{1/4}$ instead of splitting into $\|\mathbf{u}\| \leqslant \sqrt{cp \log p}$ and $\|\mathbf{u}\| > \sqrt{cp \log p}$ in (2.21) and proceed similarly to show that the normal approximation in Theorem 2.3 holds. The details are however omitted.

For a prior π, the posterior density of $\boldsymbol{\theta}$ given the observations $\mathbf{x}_1, ..., \mathbf{x}_n$, is given by

$$\pi_n(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \exp[n(\bar{\mathbf{x}}^T \boldsymbol{\theta} - \psi(\boldsymbol{\theta}))]. \tag{2.5}$$

Put $\mathbf{u} = \sqrt{n}\, \mathbf{J}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ so that $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + n^{-1/2} \mathbf{J}^{-1} \mathbf{u}$. The likelihood ratio, as a function of $\mathbf{u}$, is given by

$$Z_n(\mathbf{u}) = \exp[\sqrt{n}\, \bar{\mathbf{x}}^T \mathbf{J}^{-1} \mathbf{u} - (\psi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{J}^{-1} \mathbf{u}) - \psi(\boldsymbol{\theta}_0))]$$

if $\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{J}^{-1} \mathbf{u} \in \Theta$ and $Z_n(\mathbf{u}) = 0$ otherwise. Thus the posterior density of $\mathbf{u}$ is given by

$$\pi_n^*(\mathbf{u}) = \frac{\pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{J}^{-1} \mathbf{u})\, Z_n(\mathbf{u})}{\int \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{J}^{-1} \mathbf{w})\, Z_n(\mathbf{w})\, d\mathbf{w}}. \tag{2.6}$$

Further, setting $\boldsymbol{\Delta}_n = \sqrt{n}\, \mathbf{J}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})$, we see that $E\boldsymbol{\Delta}_n = \mathbf{0}$ and $E(\boldsymbol{\Delta}_n \boldsymbol{\Delta}_n^T) = \mathbf{I}_p$, the identity matrix of order p. Hence $E(\|\boldsymbol{\Delta}_n\|^2) = E(\mathrm{tr}(\boldsymbol{\Delta}_n \boldsymbol{\Delta}_n^T)) = p$. It then easily follows from Chebyshev's inequality that

$$\|\boldsymbol{\Delta}_n\| = O_p(\sqrt{p}). \tag{2.7}$$

Below, we present a result on the consistency of the MLE. Here, unlike Theorem 2.1 of Portnoy [21], we use a different distance measure and do not assume (2.4) of Portnoy [21]. It may be noted that (2.4) of Portnoy [21] fails to hold if the minimum eigenvalue of $\mathbf{F}$ tends to zero [e.g., multinomial distribution, see Section 3]. The result will be used in the proof of the main theorem.

For $\boldsymbol{\theta} \in \Theta$, define $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_0 = \|\mathbf{J}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|$. Observe that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_0$ does not depend on the choice of the square root $\mathbf{J}$ of $\mathbf{F}$ and is a weighted Euclidean distance of $\boldsymbol{\theta}$ from $\boldsymbol{\theta}_0$ with $\boldsymbol{\theta}_0$ as a preferred point [Critchley *et al.* [4]]. Since $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|$ has the same local behaviour as the Riemannian metric based on the Fisher information, this distance measure is arguably more intrinsic to the given statistical problem. The gain is also immediate as we can then avoid conditions (2.4) of Portnoy [21] as well as the use of Theorem 4.1 of Portnoy [21], which requires a bound, free from the dimension, on the sixth moment of the components of standardized variable $\mathbf{J}^{-1}(\mathbf{x} - \boldsymbol{\mu})$. If desired, consistency in terms of the Euclidean

distance can also be readily obtained, though with a different rate. The difference between the choice of the two distances, however, essentially disappears in fixed dimension.

THEOREM 2.1. *Assume that for all* $c>0$, $\sqrt{p/n}\, B_{1n}(c)\to 0$ *and* $p\,\|\mathbf{F}^{-1}\|/n\to 0$ *as* $n\to\infty$. *Then the MLE* $\hat{\boldsymbol{\theta}}$ *of* $\boldsymbol{\theta}$ *satisfies*

$$\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_0=\|\mathbf{J}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|=O_p(\sqrt{p/n}) \tag{2.8}$$

and so

$$\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|=O_p(\sqrt{p\,\|\mathbf{F}^{-1}\|/n})=o_p(1).$$

Proof. The proof of Theorem 2.1 of Portnoy [21] essentially carries over. Observe that $\sqrt{n}\,\mathbf{J}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ is the unique root of the equation $G(\mathbf{u})=\mathbf{0}$, where $G(\mathbf{u})=\mathbf{J}^{-1}[\psi'(\boldsymbol{\theta}_0+n^{-1/2}\mathbf{J}^{-1}\mathbf{u})-\bar{\mathbf{x}}]$. By (2.7), for $\varepsilon>0$, find $K>0$ such that $P\{\|\mathbf{J}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\|>K\sqrt{p/n}\}<\varepsilon$ and choose $c>K^2$. Following Portnoy's [21] arguments, it can now be shown that a $\|\sqrt{n}\,\mathbf{J}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|\leqslant\sqrt{cp}$ with probability larger than $1-\varepsilon$.

In a similar manner, we can restate Theorem 3.1 of Portnoy [21] on asymptotic normality in the following way. The last part of the result will be used in Theorem 2.4. The proof is omitted.

THEOREM 2.2. *Assume that for all* $c>0$, $pB_{1n}(c)/\sqrt{n}\to 0$ *and* $p\,\|\mathbf{F}^{-1}\|/n\to 0$ *as* $n\to\infty$. *Then for any vector* $\mathbf{a}$ *with* $\|\mathbf{a}\|=1$, *we have*

$$\sqrt{n}\,\mathbf{a}^T\mathbf{J}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)-\sqrt{n}\,\mathbf{a}^T\mathbf{J}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})=o_p(1) \tag{2.9}$$

and

$$\sqrt{n}\,\mathbf{a}^T\mathbf{J}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\to_d N(0,1). \tag{2.10}$$

Moreover, if for all $c\geqslant 0$, $p^2B_{2n}(c)/n\to 0$, *then*

$$\|\sqrt{n}\,\mathbf{J}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)-\sqrt{n}\,\mathbf{J}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\|=o_p(1). \tag{2.11}$$

The following is the main result of this paper.

THEOREM 2.3. *Under Conditions* (P) *and* (R), *we have*

$$\int|\pi_n^*(\mathbf{u})-\phi_p(\mathbf{u};\boldsymbol{\Delta}_n,\mathbf{I}_p)|\,d\mathbf{u}\to_p 0, \tag{2.12}$$

where $\phi_p(\cdot;\mathbf{v},\boldsymbol{\Sigma})$ *stands for the density of* $N_p(\mathbf{v},\boldsymbol{\Sigma})$.

Since the L_1-distance between two densities is the same as the total variation distance between the corresponding probabilities up to a factor of 2, (2.12) means that posterior probabilities of sets are uniformly approximated by the corresponding normal probabilities, i.e.,

$$\sup \{ |\Pr(\boldsymbol{\theta} \in B \mid X_1, ..., X_n) - \Pr(\hat{\boldsymbol{\theta}} + n^{-1/2} \mathbf{J}^{-1} \boldsymbol{\xi} \in B)| : B \in \mathscr{B}^p \} \to_p 0,$$

where $\boldsymbol{\xi}$ has $N_p(\boldsymbol{\Delta}_n, \mathbf{I}_p)$ distribution and $\mathscr{B}^p$ is the Borel σ-field on $\mathbb{R}^p$. Also, since the Hellinger distance $H(f, g) = (\int (f^{1/2} - g^{1/2})^2)^{1/2}$ satisfies

$$H^2(f, g) \leqslant \int |f - g| \leqslant \sqrt{2}\, H(f, g)$$

for any two densities f and g, the normal approximation in (2.12) holds in the sense of the Hellinger distance as well.

To prove Theorem 2.3, we use the following Lemmas 2.1–2.5. Proof of these lemmas are deferred to the appendix.

We set $\tilde{Z}_n(\mathbf{u}) = \exp[\mathbf{u}^T \boldsymbol{\Delta}_n - \frac{1}{2} \|\mathbf{u}\|^2]$ throughout the paper.

LEMMA 2.1. *For all* $\mathbf{u}$ *with* $\|\mathbf{u}\|^2 \leqslant cp \log p$, *we have*

$$|\log Z_n(\mathbf{u}) - \log \tilde{Z}_n(\mathbf{u})| \leqslant \lambda_n(c)\, \|\mathbf{u}\|^2 \tag{2.13}$$

and

$$\log Z_n(\mathbf{u}) \leqslant \mathbf{u}^T \boldsymbol{\Delta}_n - \tfrac{1}{2} \|\mathbf{u}\|^2 (1 - 2\lambda_n(c)), \tag{2.14}$$

where $\lambda_n(c) = (\sqrt{(cp \log p)/n}\, B_{1n}(0) + ((cp \log p)/n)\, B_{2n}(c \log p))/6$.

LEMMA 2.2. *With probability tending to one,*

$$\log Z_n(\mathbf{u}) \leqslant -\tfrac{1}{4} cp \log p \qquad \text{on} \quad \|\mathbf{u}\|^2 > cp \log p. \tag{2.15}$$

LEMMA 2.3. *For any* $c > 0$, *we have*

$$\left(\int \tilde{Z}_n(\mathbf{u})\, d\mathbf{u} \right)^{-1} \int_{\|\mathbf{u}\|^2 \leqslant cp \log p} |Z_n(\mathbf{u}) - \tilde{Z}_n(\mathbf{u})|\, d\mathbf{u} \to_p 0. \tag{2.16}$$

LEMMA 2.4. *For any* m *and* $\varepsilon > 0$, *we can find* $c > 0$ *such that with probability greater than* $1 - \varepsilon$,

$$\left(\int \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{u})\, d\mathbf{u} \right)^{-1} \int_{\|\mathbf{u}\|^2 > cp \log p} \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{J}^{-1} \mathbf{u})\, Z_n(\mathbf{u})\, d\mathbf{u}$$
$$\leqslant \exp[-mp \log p]. \tag{2.17}$$

LEMMA 2.5. *Given an $m>0$ and $\varepsilon>0$, a constant $c>0$ can be found so that with probability greater than $1-\varepsilon$,*

$$\int_{\|\mathbf{u}\|^2>cp} \phi_p(\mathbf{u};\, \mathbf{\Delta}_n,\, \mathbf{I}_p)\, d\mathbf{u} \leqslant e^{-mp}. \tag{2.18}$$

Proof of Theorem 2.3. By (2.6) and the definition of $\tilde{Z}_n(\mathbf{u})$, we have

$$\begin{aligned}
&\int |\pi_n^*(\mathbf{u}) - \phi_p(\mathbf{u};\, \mathbf{\Delta}_n,\, \mathbf{I}_p)|\, d\mathbf{u} \\
&\quad = \int \left| \frac{Z_n(\mathbf{u})\, \pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u})}{\int Z_n(\mathbf{w})\, \pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{w})\, d\mathbf{w}} \right. \\
&\qquad \left. - \frac{\pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{u})}{\int \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{w})\, d\mathbf{w}} \right| d\mathbf{u}.
\end{aligned} \tag{2.19}$$

By adding and subtracting the term

$$\frac{Z_n(\mathbf{u})\, \pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u})}{\int \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{w})\, d\mathbf{w}}$$

inside the modulus on the right hand side (RHS) of (2.19) and using the triangle inequality, we can bound the RHS of (2.19) by

$$\begin{aligned}
&\left| \left(\int Z_n(\mathbf{w})\, \pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{w})\, d\mathbf{w} \right)^{-1} - \left(\int \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{w})\, d\mathbf{w} \right)^{-1} \right| \\
&\qquad \times \int Z_n(\mathbf{u})\, \pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u})\, d\mathbf{u} \\
&\qquad + \left(\int \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{w})\, d\mathbf{w} \right)^{-1} \\
&\qquad \times \int |\pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u})\, Z_n(\mathbf{u}) - \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{u})|\, d\mathbf{u}.
\end{aligned} \tag{2.20}$$

The first term in (2.20) is equal to

$$\begin{aligned}
&\left(\int \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{w})\, d\mathbf{w} \right)^{-1} \\
&\qquad \times \left| \int \pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u})\, Z_n(\mathbf{u})\, d\mathbf{u} - \int \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{u})\, d\mathbf{u} \right|,
\end{aligned}$$

which is clearly dominated by the second term in (2.20). Therefore, it suffices to bound the latter. To this end, we split the integral in the numerator in regions $\|\mathbf{u}\|^2 \leqslant cp \log p$ and $\|\mathbf{u}\|^2 > cp \log p$, where c is to be chosen later, and estimate the difference by the sum of the integrands on the latter region to obtain the bound

$$\left(\int \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{w})\, d\mathbf{w}\right)^{-1}$$

$$\times \int_{\|\mathbf{u}\|^2 \leqslant cp \log p} |\pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u})\, Z_n(\mathbf{u}) - \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{u})|\, d\mathbf{u}$$

$$+ \left(\int \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{w})\, d\mathbf{w}\right)^{-1}$$

$$\times \int_{\|\mathbf{u}\|^2 > cp \log p} Z_n(\mathbf{u})\, \pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u})\, d\mathbf{u}$$

$$+ \int_{\|\mathbf{u}\|^2 > cp \log p} \phi_p(\mathbf{u}; \boldsymbol{\Delta}_n, \mathbf{I}_p)\, d\mathbf{u}. \tag{2.21}$$

Using Lemmas 2.4 and 2.5 respectively, the last two terms in (2.21) can be made as small as we please with probability arbitrarily close to one by choosing c large enough. For this chosen c, let F denote the set $\{\mathbf{u}: \|\mathbf{u}\|^2 \leqslant cp \log p\}$. Then the first term on the RHS of (2.21) is dominated by

$$\sup_{\mathbf{u} \in F} \left| \frac{\pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u})}{\pi(\boldsymbol{\theta}_0)} - 1 \right| \frac{\int_F Z_n(\mathbf{u}\, d\mathbf{u})}{\int \tilde{Z}_n(\mathbf{u})\, d\mathbf{u}}$$

$$+ \left(\int \tilde{Z}_n(\mathbf{u})\, d\mathbf{u}\right)^{-1} \int_F |Z_n(\mathbf{u}) - \tilde{Z}_n(\mathbf{u})|\, d\mathbf{u}. \tag{2.22}$$

Since $|e^x - 1| \leqslant |x|\, e^{|x|} \leqslant 2\,|x|$ for sufficiently small $|x|$ and

$$\sup \left\{ \left| \log \frac{\pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u})}{\pi(\boldsymbol{\theta}_0)} \right| : \mathbf{u} \in F \right\}$$

$$\leqslant 2K_n(c) \sqrt{\|\mathbf{F}^{-1}\|\, cp(\log p)/n} \to 0$$

by (2.3), it follows that

$$\sup_{\mathbf{u} \in F} \left| \frac{\pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u})}{\pi(\boldsymbol{\theta}_0)} - 1 \right| \to 0. \tag{2.23}$$

The last term in (2.22) converges to zero by Lemma 2.3. This, in particular, implies that $\int_F Z_n(\mathbf{u})\,d\mathbf{u}/\int \tilde{Z}_n(\mathbf{u})\,d\mathbf{u}$ remains bounded in probability as

$$\frac{\int_F Z_n(\mathbf{u})\,d\mathbf{u}}{\int \tilde{Z}_n(\mathbf{u})\,d\mathbf{u}} \leqslant 1 + \left(\int \tilde{Z}_n(\mathbf{u})\,d\mathbf{u}\right)^{-1} \int_F |Z_n(\mathbf{u}) - \tilde{Z}_n(\mathbf{u})|\,d\mathbf{u}.$$

Hence the expression in (2.22) is arbitrarily small with probability arbitrarily close to unity, proving the theorem.

From Theorem 2.3, we easily obtain the consistency of the posterior distribution.

COROLLARY 2.1. *Under the conditions of Theorem* 2.3, *the posterior probability of* $\{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta\}$ *for any fixed* $\delta > 0$ *converges to one in probability. In fact, if* $p \to \infty$, *there is a* $c > 0$ *such that the posterior probability of* $\{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leqslant \sqrt{cp}\,\|\mathbf{F}^{-1}\|/n\}$ *converges to* 1 *in probability.*

To prove the corollary, note that by Theorem 2.3 the posterior probability of $\{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta\}$ is approximated by $\Pr\{n^{-1/2}\,\|\mathbf{J}^{-1}\boldsymbol{\xi}\| < \delta\}$ in probability, where $\boldsymbol{\xi}$ has distribution $N_p(\boldsymbol{\Delta}_n, \mathbf{I}_p)$ and hence it suffices to show that the latter converges to 1 in probability. Now

$$n^{-1/2}\,\|\mathbf{J}^{-1}\boldsymbol{\Delta}_n\| \leqslant \|\boldsymbol{\Delta}_n\|\,\sqrt{\|\mathbf{F}^{-1}\|/n} = O_p(\sqrt{p\,\|\mathbf{F}^{-1}\|/n}) = o_p(1),$$

so that on a set whose probability tends to one,

$$\begin{aligned}\Pr\{n^{-1/2}\,\|\mathbf{J}^{-1}\boldsymbol{\xi}\| \geqslant \delta\} &\leqslant \Pr\left\{n^{-1/2}\,\|\mathbf{J}^{-1}\|\,\|\boldsymbol{\xi} - \boldsymbol{\Delta}_n\| \geqslant \frac{\delta}{2}\right\} \\ &= \Pr\left(n^{-1}\,\|\mathbf{F}^{-1}\|\,Y \geqslant \frac{\delta^2}{4}\right), \qquad (2.24)\end{aligned}$$

where Y has a central chi-square distribution with p degrees of freedom. As $E(n^{-1}\,\|\mathbf{F}^{-1}\|\,Y) = p\,\|\mathbf{F}^{-1}\|/n \to 0$, the RHS of (2.24) tends to 0, proving the first part of the corollary. For the second part, we proceed similarly and with a sufficiently large c, end up with the bound $\Pr\{Y > cp/4\}$ on the RHS of (2.24). The result now follows by a simple large deviation estimate; see Bahadur [1].

Remark 2.1. If in Condition (R), we strengthen

$$B_{1n}(0)\,p^{3/2}(\log p)^{1/2}/\sqrt{n} \to 0,$$

$$B_{2n}(c \log p)\,p^2(\log p)/n \to 0$$

$$K_n(c)\,\sqrt{\|\mathbf{F}^{-1}\|\,p(\log p)/n} \to 0$$

to

$$B_{1n}(0)\ p^2(\log p)/\sqrt{n} \to 0,$$
$$B_{2n}(c \log p)\ p^{5/2}(\log p)^{3/2}/n \to 0$$
$$K_n(c)\ \sqrt{\|\mathbf{F}^{-1}\|\ p^{3/2}(\log p)^{3/2}/n} \to 0$$

respectively, then

$$\int \|\mathbf{u}\|\ |\pi_n^*(\mathbf{u}) - \phi_p(\mathbf{u}; \boldsymbol{\Delta}_n, \mathbf{I}_p)|\ d\mathbf{u} \to_p 0, \tag{2.25}$$

which yields the following asymptotic representation of the posterior mean $\tilde{\boldsymbol{\theta}}$:

$$\sqrt{n}\ \mathbf{J}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \boldsymbol{\Delta}_n + \mathbf{o}_p(1). \tag{2.26}$$

Thus the posterior mean is asymptotically normal and asymptotically efficient.

Theorem 2.3 is a result of theoretical nature. It is itself not very useful for the actual approximation of the posterior since the approximation is dependent on $\boldsymbol{\Delta}_n$ and $\mathbf{J}$, which involve the unknown value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$. We now present a version of Theorem 2.3 which replaces the unknown parameter by its estimate.

THEOREM 2.4. *Assume Condition* (P) *and Condition* (R) *and suppose that*

$$\log\det \mathbf{F}(\boldsymbol{\theta}) - \log\det \mathbf{F}(\boldsymbol{\theta}_0) \to 0 \qquad \textit{and} \qquad \operatorname{tr}((\mathbf{F}(\boldsymbol{\theta}))^{-1}\, \mathbf{F}(\boldsymbol{\theta}_0)) - p \to 0$$

uniformly on $\{\boldsymbol{\theta}\colon \|\mathbf{J}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 \leqslant cp/n\}$. *Let* $\mathbf{v} = \sqrt{n}\ \hat{\mathbf{J}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, *where* $\hat{\boldsymbol{\theta}}$ *is the MLE and* $\hat{\mathbf{J}}$ *is a square root of the covariance matrix evaluated at* $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. *Then the posterior density* $\hat{\pi}_n(\mathbf{v})$ *of* $\mathbf{v}$ *is approximately standard normal in the sense that*

$$\int |\hat{\pi}_n(\mathbf{v}) - \phi_p(\mathbf{v}; \mathbf{0}, \mathbf{I}_p)|\ d\mathbf{v} \to_p 0. \tag{2.27}$$

Proof. Put $\mathbf{w} = \sqrt{n}\ \mathbf{J}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$. By Theorem 2.3 and the invariance of the L_1-distance under a change of location, the posterior density $\tilde{\pi}_n(\mathbf{w})$ of $\mathbf{w}$ satisfies

$$\int |\tilde{\pi}_n(\mathbf{w}) - \phi_p(\mathbf{w}; \boldsymbol{\Delta}_n - \sqrt{n}\ \mathbf{J}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \mathbf{I}_p)|\ d\mathbf{w} \to_p 0. \tag{2.28}$$

We now show that the normal density appearing in (2.28) can be approximated by the standard normal density in the L_1-distance in probability. It suffices to bound their entropy distance, which is equal to $\frac{1}{2}\|\Delta_n - \sqrt{n}\,\mathbf{J}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|^2$ and so converges to 0 in probability, by the last part of Theorem 2.2. Thus by the invariance of the L_1-distance under a change of scale, we have

$$\int |\hat{\pi}_n(\mathbf{v}) - \phi_p(\mathbf{v}; \mathbf{0}, \hat{\mathbf{J}}\mathbf{F}^{-1}\hat{\mathbf{J}})|\, d\mathbf{v} \to_p 0. \tag{2.29}$$

The entropy distance between the normal density appearing in (2.29) and the standard normal is

$$\tfrac{1}{2}\operatorname{tr}(\hat{\mathbf{J}}^{-1}\mathbf{F}\hat{\mathbf{J}}^{-1} - \mathbf{I}_p) + \tfrac{1}{2}\log\det(\hat{\mathbf{J}}^{-1}\mathbf{F}\hat{\mathbf{J}}^{-1} - \mathbf{I}_p),$$

which converges to zero under the additional assumptions made above. This completes the proof.

Remark 2.2. Theorem 2.4 readily yields approximate highest posterior density regions for $\boldsymbol{\theta}$ which are ellipsoids centered at the MLE.

3. APPLICATION TO THE MULTINOMIAL DISTRIBUTION AND BAYESIAN DENSITY ESTIMATION

Consider the multinomial distribution with $(p+1)$ cells. Let $\mathbf{x} = (x_1, ..., x_p)$ stands for the vector of observations and $\boldsymbol{\pi} = (\pi_1, ..., \pi_p)$ stand for the vector of probabilities of the cells excepting the zeroth one. The natural parameter is given by $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$, where $\theta_j = \log(\pi_j/(1 - \sum_{k=1}^p \pi_k))$. We assume that the true value of θ_j's are bounded, so π_j's are of the order p^{-1}. Note that $\boldsymbol{\pi}$ is the mean vector and variance-covariance matrix or the Fisher information is given by $\mathbf{F} = \mathbf{D} - \boldsymbol{\pi}\boldsymbol{\pi}^T$, where $\mathbf{D} = \operatorname{diag}(\pi_1, ..., \pi_p)$. Thus

$$\mathbf{F}^{-1} = \mathbf{D}^{-1} + \frac{\mathbf{D}^{-1}\boldsymbol{\pi}\boldsymbol{\pi}^T\mathbf{D}^{-1}}{1 - \boldsymbol{\pi}^T\mathbf{D}^{-1}\boldsymbol{\pi}} = \mathbf{D}^{-1} + \frac{\mathbf{1}\mathbf{1}^T}{1 - \boldsymbol{\pi}^T\mathbf{D}^{-1}\boldsymbol{\pi}}, \tag{3.1}$$

where $\mathbf{1}$ is the p-vector with all entries equal to one. Note that $1 - \boldsymbol{\pi}^T\mathbf{D}^{-1}\boldsymbol{\pi} = 1 - \sum_{j=1}^p \pi_j = \pi_0$ (say) is also of the order p^{-1}. Thus $\|\mathbf{F}^{-1}\| \leqslant \operatorname{tr}(\mathbf{F}^{-1}) = O(p^2)$. In general, this rate cannot be improved, since for the case $\boldsymbol{\theta} = \mathbf{0}$ (i.e., all π_j's are $(p+1)^{-1}$), the largest eigenvalue of $\mathbf{F}^{-1}$ is of the order p^2. Also, $\operatorname{tr}(\mathbf{F}) \leqslant 1$. It can be verified that

$$\mathbf{J} = \mathbf{D}^{1/2} - \frac{\boldsymbol{\pi}\boldsymbol{\pi}^T\mathbf{D}^{-1/2}}{1 + \sqrt{1 - \boldsymbol{\pi}^T\mathbf{D}^{-1}\boldsymbol{\pi}}} \tag{3.2}$$

and

$$\mathbf{J}^{-1} = \mathbf{D}^{-1/2} + \frac{\mathbf{D}^{-1}\boldsymbol{\pi}\boldsymbol{\pi}^T\mathbf{D}^{-1/2}}{1 - \boldsymbol{\pi}^T\mathbf{D}^{-1}\boldsymbol{\pi} + \sqrt{1 - \boldsymbol{\pi}^T\mathbf{D}^{-1}\boldsymbol{\pi}}} \tag{3.3}$$

are square roots of $\mathbf{F}$ and $\mathbf{F}^{-1}$ respectively. Let $\mathbf{a}$ be a unit p-vector. We need to calculate the third and fourth order absolute moments, with $\boldsymbol{\theta}$ as the underlying parameter, of

$$\mathbf{a}^T\mathbf{J}^{-1}(\mathbf{x} - \boldsymbol{\pi}) = \sum_{j=1}^{p} a_j \pi_j^{-1/2}(x_j - \pi_j) + \alpha\left(\sum_{j=1}^{p} a_j\right)\sum_{j=1}^{p} \pi_j^{1/2}(x_j - \pi_j),$$

where $\alpha^{-1} = \pi_0 + \sqrt{\pi_0}$. Using the facts that π_j's are of the order p^{-1}, $\alpha = O(\sqrt{p})$ and $|\sum_{j=1}^{p} a_j| \leqslant \sqrt{p}$, it can be verified that the third moment is at most of the order $p^{3/2}$ and the fourth moment is at most of the order p^2. The order remains the same even if $\boldsymbol{\theta}$ is replaced by some $\boldsymbol{\theta}^*$ satisfying $\|\mathbf{J}(\boldsymbol{\theta}^* - \boldsymbol{\theta})\|^2 \leqslant cp(\log p)/n$. To see that, first note that $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\| \leqslant \|\mathbf{J}^{-1}\|\,\|\mathbf{J}(\boldsymbol{\theta}^* - \boldsymbol{\theta})\| = O(p^{3/2}(\log p)^{1/2}/\sqrt{n})$, so the components of $\boldsymbol{\theta}^*$ are again uniformly bounded if $p^3(\log p)/n \to 0$, and hence the corresponding cell probabilities are exactly of order p^{-1}. Similar calculations will show that the order of the third and fourth absolute moments remain $p^{3/2}$ and p^2 respectively, even if $\boldsymbol{\theta}_0$ is replaced by $\boldsymbol{\theta}$. Thus under the condition $p^6(\log p)/n \to 0$, Condition (R) verifies provided the constants $K_n(c)$'s do not grow faster than $p^{3/2}$. Apart from the priors for which θ_j's are independently distributed, the conjugate prior

$$\frac{\Gamma(\alpha_0 + \alpha_1 + \cdots + \alpha_p)}{\Gamma(\alpha_0)\,\Gamma(\alpha_1)\cdots\Gamma(\alpha_p)} \exp\left(\sum_{j=1}^{p} \alpha_j\theta_j\right)\left(1 + \sum_{j=1}^{p} e^{\theta_j}\right)^{-(\alpha_0 + \alpha_1 + \cdots + \alpha_p)}$$

also satisfies Condition (P) with $K_n(c) = O(\sqrt{p})$, provided $\{\alpha_j\}$ and $\{\alpha_j^{-1}\}$ are bounded sequences. Finally, the condition $p^4/n \to 0$ suffices for the consistency of the MLE.

To verify the two additional conditions of Theorem 2.4, let π_j and π_{0j}, $j = 0, 1, ..., p$, denote the cell probabilities corresponding to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ respectively, where $\|\mathbf{J}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 \leqslant cp/n$. As mentioned above, π_j and π_{0j}, $j = 0, 1, ..., p$, are of the order p^{-1}. Using the form (3.1) of the inverse information matrix, straightforward computations show that

$$\operatorname{tr}(\mathbf{F}^{-1}(\boldsymbol{\theta})\,\mathbf{F}(\boldsymbol{\theta}_0)) = \sum_{j=0}^{p} \frac{\pi_{0j}(1 - \pi_{0j})}{\pi_j}.$$

Also, $\det \mathbf{F}(\boldsymbol{\theta}) = \prod_{j=0}^{p} \pi_j$. Below, C will stand for a generic constant. Now

$$\begin{aligned}
|\pi_j - \pi_{0j}| &= \left| \frac{e^{\theta_j}}{1+\sum_{l=1}^{p} e^{\theta_l}} - \frac{e^{\theta_{0j}}}{1+\sum_{l=1}^{p} e^{\theta_{0l}}} \right| \\
&\leqslant \frac{|e^{\theta_j} - e^{\theta_{0j}}|}{1+\sum_{l=1}^{p} e^{\theta_l}} + e^{\theta_{0j}} \frac{|\sum_{l=1}^{p} e^{\theta_l} - \sum_{l=1}^{p} e^{\theta_{0l}}|}{(1+\sum_{l=1}^{p} e^{\theta_l})(1+\sum_{l=1}^{p} e^{\theta_{0l}})} \\
&\leqslant Cp^{-1} |\theta_j - \theta_{0j}| + Cp^{-2} \sum_{l=1}^{p} |\theta_l - \theta_{0l}| \\
&\leqslant Cp^{-1} |\theta_j - \theta_{0j}| + Cp^{-3/2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|. \qquad (3.4)
\end{aligned}$$

Thus

$$\begin{aligned}
|\mathrm{tr}(\mathbf{F}^{-1}(\boldsymbol{\theta})\, \mathbf{F}(\boldsymbol{\theta}_0)) - p| &= \left| \sum_{j=0}^{p} \frac{\pi_{0j}(1-\pi_{0j}) - \pi_j(1-\pi_j)}{\pi_j} \right| \\
&\leqslant Cp^2 \max_{0 \leqslant j \leqslant p} |\pi_j - \pi_{0j}| \\
&\leqslant Cp \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \\
&\leqslant Cp^2 \|\mathbf{J}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\| \\
&= O(p^{5/2}/\sqrt{n})
\end{aligned}$$

and

$$\begin{aligned}
|\log \det \mathbf{F}(\boldsymbol{\theta}) - \log \det \mathbf{F}(\boldsymbol{\theta}_0)| &= \left| \sum_{j=0}^{p} (\log \pi_j - \log \pi_{0j}) \right| \\
&\leqslant Cp^2 \max_{0 \leqslant j \leqslant p} |\pi_j - \pi_{0j}| \\
&= O(p^{5/2}/\sqrt{n}).
\end{aligned}$$

Thus the conditions hold if $p^5/n \to 0$.

The result on the consistency of the posterior distribution for the multinomial distribution has an interesting link with a Bayesian density estimation problem. Suppose we have a positive Lipschitz continuous density f on the unit interval which we wish to estimate using a Bayesian method. We observe samples $y_1, y_2, ..., y_n$ from f. Our prior will be supported on certain histograms. Depending on n, choose an integer $p = p_n$ so that $p \to \infty$ and $p^6(\log p)/n \to 0$. Now divide the unit interval into the $(p+1)$ subintervals of length $1/(p+1)$, to be denoted by $\Delta_0, \Delta_1, ..., \Delta_p$. Define $\pi_0, \pi_1, ..., \pi_p$ to be the probabilities of the subintervals under the density f. Under the model, the vector of indicators $\mathbf{x}_i = (x_{i0}, ..., x_{ip})$, $i = 1, ..., n$, are sufficient for the data $y_1, ..., y_n$ and are i.i.d. multinomial with $(p+1)$ cells and probabilities $\pi_0, ..., \pi_p$. Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ be the natural parameter. Let the set of all histograms defined by the partition $\{\Delta_0, \Delta_1, ..., \Delta_p\}$ be denoted by

$\mathscr{F}_n$. This $\mathscr{F}_n$ can be thought of as a sieve in the sense of Grenander [12] and possesses the following approximation property: Let f_n be the density defined by $f_n(x) = (p+1)\sum_{j=0}^{p} \pi_j I\{x \in \Delta_j\}$. Then $f_n \in \mathscr{F}_n$ and $\int (f_n(x) - f(x))^2\, dx = O(p^{-2})$ as $n \to \infty$. When the sample size is n, we put a prior Π_n on f by defining a prior on $\mathscr{F}_n$ through a prior density on θ_j's satisfying the required condition of the above discussion. For example, θ_j's could be independently distributed or could have a Dirichlet distribution. Thus we have a simple sequence of priors for which posterior could easily be calculated, particularly if the prior is Dirichlet. Let the true value of f be f_0 and the corresponding f_n, π_j, θ_j and $\boldsymbol{\theta}$ be denoted by f_{0n}, π_{0j}'s, θ_{0j}'s and $\boldsymbol{\theta}_0$ respectively. We shall show that the posterior for f concentrates near f_0 at a certain rate.

With f_{0n} as defined above and every f in the support of the prior Π_n, note that f and f_{0n} are constant on each Δ_j taking values π_j and π_{0j} respectively, $j = 0, 1, ..., p$. Hence by the definition of f_{0n}

$$\int (f(x) - f_{0n}(x))(f_{0n}(x) - f_0(x))\, dx$$

$$= \sum_{j=0}^{p} (\pi_j - \pi_{0j}) \int_{\Delta_j} (f_{0n}(x) - f_0(x))\, dx = 0.$$

Therefore

$$\int (f(x) - f_0(x))^2\, dx = \int (f(x) - f_{0n}(x))^2\, dx + \int (f_{0n}(x) - f_0(x))^2\, dx.$$

The second term on the right hand side of the last display is non-stochastic and converges to 0 at the rate p^{-2}. Note that, since the true density f_0 does not belong to the support of the prior, the density f_{0n}, which is the density closest to f_0 in the support of the prior, works as a proxy for the true f_0.

The first term on the RHS of the last display is equal to $(p+1)\sum_{j=0}^{p} (\pi_j - \pi_{0j})^2$. We shall show that, for a sufficiently large c, posterior probability of the set

$$\left\{\boldsymbol{\theta}: (p+1) \sum_{j=0}^{p} (\pi_j - \pi_{0j})^2 < cp^2/n\right\}$$

converges to 1 in probability. The true θ_{0j}'s lie in a compact interval $[a, b]$ independent of n. This follows by positivity and continuity of f_0. Therefore if $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta$, $\delta > 0$, θ_j's lie in a slightly bigger interval $[a - \delta, b + \delta]$. As argued in (3.4), on $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta$,

$$|\pi_j - \pi_{0j}| \leqslant Cp^{-1} |\theta_j - \theta_{0j}| + Cp^{-3/2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|, \qquad j = 0, ..., p.$$

Squaring and adding, it follows that on $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\delta$,

$$(p+1)\sum_{j=0}^{p}(\pi_j-\pi_{0j})^2<Cp^{-1}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2.$$

By posterior consistency, posterior probability of $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\delta$ converges to one. Moreover, for a large enough constant c, posterior probability of $\{\boldsymbol{\theta}: \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 \leqslant cp\|\mathbf{F}^{-1}\|/n\}$ converges to 1, where $\mathbf{F}$ is, as above, the Fisher information at $\boldsymbol{\theta}_0$. Since $\|\mathbf{F}^{-1}\|=O(p^2)$, the claim follows. When $p^6(\log p)/n\to 0$, the bias (f_n-f_0) contributes more to the error than the variability $(f-f_n)$. Choosing $p=n^{1/6}/(\log n)^{(1/6)+\varepsilon}$, $\varepsilon>0$, we see that for a sufficiently large constant c,

$$\Pr\left\{f: \int (f(x)-f_0(x))^2\,dx \leqslant cn^{-1/3}(\log n)^{(1/3)+2\varepsilon} \mid y_1, \ldots, y_n\right\}\to_p 0.$$

Gasparini [6], like us, considered priors supported on histograms where the window length was also given a prior and the mass was distributed to the intervals according to a Dirichlet process on natural numbers. He showed consistency of the Bayes estimate of the density for weak and variation neighbourhoods. Ghosal *et al.* [9] considered Dirichlet mixtures of normals as a prior on the densities and established weak and strong consistencies of the posterior distribution. Rates of convergence of posterior distribution are discussed only recently by Ghosal *et al.* [11]. If the densities belong to the Hölder class of order α (see Example 1 of Wong and Shen [22]), they constructed priors based on bracketings or splines that achieve the optimal rate $n^{-\alpha/(2\alpha+1)}$ of convergence of the posterior distribution for the Hellinger distance. For the special case $\alpha=1$, the Hölder class essentially reduces to the class of Lipschitz continuous densities and the convergence rate $n^{-1/3}$ is obtained. Although priors constructed by Ghosal *et al.* [11] lead to a better rate of convergence of the posterior, computation of the posterior for those priors is much more involved. On the other hand, for the histogram type prior constructed above, the posterior computation is much simpler. The normal approximation established in Theorem 2.4 may also be used to simplify computations further.

4. APPLICATION TO THE ESTIMATION OF THE MEAN OF AN INFINITE DIMENSIONAL NORMAL DISTRIBUTION

Suppose we observe n i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from an infinite dimensional normal population with mean $\boldsymbol{\theta}=(\theta_1, \theta_2, \ldots)$ and the identity operator on $\ell_2=\{(y_1, y_2, \ldots): \sum_{i=1}^{\infty} y_i^2<\infty\}$ as the covariance, i.e.,

components of each observations are also independent. It is assumed that $\boldsymbol{\theta} \in \ell_2$. We shall use the ℓ_2-norm to measure distances. The rate at which $\boldsymbol{\theta}$ may be estimated depends on rate at which θ_i decays to 0. Pinsker [17] showed that on the ellipsoid $\{\boldsymbol{\theta}: \sum_{i=1}^{\infty} i^{2q}\theta_i^2 \leqslant Q\}$, the minimax rate of convergence is $n^{-q/(2q+1)}$. Diaconis and Freedman [5] showed that normal approximation to the posterior distribution does not hold for this model in the usual sense. By explicit computations, Zhao [23] showed that for a suitable normal prior, the posterior mean converges at the minimax rate. In the following, we show, by the results of Section 2, that for a general class of a sequence of priors which are not necessarily normal, the posterior distribution also converges at the rate $n^{-q/(2q+1)}$.

At stage n, a prior Π_n for $\boldsymbol{\theta}$ is obtained by putting a prior on its first p components, where $p = p_n \to \infty$, and assigning the rest to 0. As the posterior depends only the first p co-ordinates, we may assume the setup of Section 2 where distributions are p-dimensional normal. In this case, since the information matrix is identity, it easy to see that B_{1n} and B_{2n} are constants in n. One may also choose a prior to satisfy (2.4). Thus Condition (R) holds if $p^3(\log p)/n \to 0$. Denoting the true mean by $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_{0,n} = (\theta_1, ..., \theta_p, 0, ...)$, we have

$$\|\boldsymbol{\theta}_{0,n} - \boldsymbol{\theta}_0\|^2 = \sum_{i=p+1}^{\infty} \theta_{i0}^2 \leqslant p^{-2q} \sum_{i=p+1}^{\infty} i^{2q}\theta_{i0}^2 = O(p^{-2q}).$$

By Corollary 2.1, for a sufficiently large c,

$$\Pi_n\{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}_{0,n}\| \leqslant c\sqrt{p/n} \mid \mathbf{x}_1, ..., \mathbf{x}_n\} \to_p 1$$

and so

$$\Pi_n\{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leqslant c \max(\sqrt{p/n}, p^{-q}) \mid \mathbf{x}_1, ..., \mathbf{x}_n\} \to_p 1.$$

The best choice of p is thus $n^{1/(2q+1)}$, for which the best possible rate $n^{-q/(2q+1)}$ is obtained. Condition (R) is satisfied for this choice of p if $q > 1$. It is interesting to note that, although the normal approximation to the posterior distribution of the infinite dimensional parameter does not hold, posterior distribution of a sequence of parametric functions that depend only $\theta_1, ..., \theta_p$ may be approximated using the normal approximation to the posterior distribution of $(\theta_1, ..., \theta_p)$, provided $p^3(\log p)/n \to 0$.

APPENDIX: PROOF OF THE LEMMAS

Proof of Lemma 2.1. Let $\mathbf{u}$ be such that $\|\mathbf{u}\|^2 \leqslant cp \log p$. We have $\log Z_n(\mathbf{u}) - \log \tilde{Z}_n(\mathbf{u}) = -(\psi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u}) - \psi(\boldsymbol{\theta}_0))$. Now by equation (2.1) of

Portnoy [21], for some $\tilde{\boldsymbol{\theta}}$ lying on the line segment joining $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u}$,

$$\begin{aligned}
&|\psi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u}) - \psi(\boldsymbol{\theta}_0)| \\
&\quad = |\tfrac{1}{6} E(\mathbf{u}^T\mathbf{V})^3 + \tfrac{1}{24}\{E_{\tilde{\theta}}(\mathbf{u}^T\mathbf{V})^4 - 3[E_{\tilde{\theta}}(\mathbf{u}^T\mathbf{V})^2]^2\}| \\
&\quad \leqslant \tfrac{1}{6}(n^{-1/2}\|\mathbf{u}\|^3 B_{1n}(0) + n^{-1}\|\mathbf{u}\|^4 B_{2n}(c\log p)) \\
&\quad \leqslant \lambda_n(c)\|\mathbf{u}\|^2. \qquad \text{(A.1)}
\end{aligned}$$

This proves (2.13) while (2.14) is an obvious consequence of (2.13).

Proof of Lemma 2.2. By the convexity of $\psi(\cdot)$, it follows that the likelihood function decreases monotonically if $\boldsymbol{\theta}$ moves away from the MLE $\hat{\boldsymbol{\theta}}$ along any line. Given any $\varepsilon > 0$, using Theorem 2.1, choose $C > 0$ large enough so that $\|\sqrt{n}\,\mathbf{J}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\| \leqslant \sqrt{Cp}$ and $\|\boldsymbol{\Delta}_n\| \leqslant \sqrt{Cp}$ with probability greater than $1-\varepsilon$. For a given $\mathbf{u}$ with $\|\mathbf{u}\| > \sqrt{cp\log p}$, let $\boldsymbol{\xi}$ be the point on the line segment joining $\mathbf{u}$ and $\sqrt{n}\,\mathbf{J}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ which satisfies $\|\boldsymbol{\xi}\|^2 = cp\log p$; such a point exists with probability greater than $1-\varepsilon$ for large n. Then by the likelihood decreasing property mentioned above and (2.14), we have for sufficiently large n, with probability greater than $1-\varepsilon$,

$$\begin{aligned}
\log Z_n(\mathbf{u}) &\leqslant \log Z_n(\boldsymbol{\xi}) \\
&\leqslant \boldsymbol{\xi}^T\boldsymbol{\Delta}_n - \tfrac{1}{2}(1 - 2\lambda_n(c))\|\boldsymbol{\xi}\|^2 \\
&\leqslant Cp - \tfrac{1}{2}(1 - 2\lambda_n(c))\, cp\log p \\
&\leqslant -\tfrac{1}{4}\, cp\log p. \qquad \text{(A.2)}
\end{aligned}$$

Proof of Lemma 2.3. Using the fact that $|e^x - e^y| \leqslant |x-y|\max\{e^x, e^y\}$ and (2.14), for $\|\mathbf{u}\|^2 \leqslant cp\log p$, we have

$$\begin{aligned}
&|Z_n(\mathbf{u}) - \tilde{Z}_n(\mathbf{u})| \\
&\quad \leqslant |\log Z_n(\mathbf{u}) - \log \tilde{Z}_n(\mathbf{u})| \exp[\mathbf{u}^T\boldsymbol{\Delta}_n - \tfrac{1}{2}(1 - 2\lambda_n(c))\|\mathbf{u}\|^2] \\
&\quad \leqslant \lambda_n(c)\|\mathbf{u}\|^2 \exp[\mathbf{u}^T\boldsymbol{\Delta}_n - \tfrac{1}{2}(1 - 2\lambda_n(c))\|\mathbf{u}\|^2], \qquad \text{(A.3)}
\end{aligned}$$

where $\lambda_n(c)$ is as defined in Lemma 2.1. Integration with respect to $\mathbf{u}$ and some manipulations yield that

$$\begin{aligned}
\int_{\|\mathbf{u}\|^2 \leqslant cp\log p} |Z_n(\mathbf{u}) - \tilde{Z}_n(\mathbf{u})|\, d\mathbf{u} &\leqslant [p + (1 - 2\lambda_n(c))^{-1}\|\boldsymbol{\Delta}_n\|^2] \\
&\quad \times \lambda_n(c)(2\pi)^{p/2}(1 - 2\lambda_n(c))^{-(p/2)+1} \\
&\quad \times \exp[(1 + 2\lambda_n(c))^{-1}\|\boldsymbol{\Delta}_n\|^2]. \qquad \text{(A.4)}
\end{aligned}$$

Since $\int \tilde{Z}_n(\mathbf{u})\, d\mathbf{u} = (2\pi)^{p/2} \exp[\frac{1}{2}\|\mathbf{\Delta}_n\|^2]$, $\|\mathbf{\Delta}_n\|^2 = O_p(p)$ and $p\lambda_n(c) \to 0$, the desired result easily follows.

Proof of Lemma 2.4. By Lemma 2.2, we have with probability tending to one, $Z_n(\mathbf{u}) \leqslant \exp[-(cp/4)\log p]$ on $\|\mathbf{u}\| > \sqrt{cp \log p}$. Now

$$\begin{aligned}
&\int_{\|\mathbf{u}\| > \sqrt{cp \log p}} \pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{J}^{-1}\mathbf{u})\, Z_n(\mathbf{u})\, d\mathbf{u} \\
&\quad \leqslant \exp\left[-\frac{1}{4} cp \log p \right] n^{p/2} (\det \mathbf{F})^{1/2} \int \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&\quad = \exp\left[-\frac{1}{4} cp \log p + \frac{p}{2} \log n + \frac{1}{2} \log \det \mathbf{F} \right]. \qquad \text{(A.5)}
\end{aligned}$$

Note that $\int \pi(\boldsymbol{\theta}_0)\, \tilde{Z}_n(\mathbf{u})\, d\mathbf{u} = \pi(\boldsymbol{\theta}_0)(2\pi)^{p/2} \exp[\|\mathbf{\Delta}_n\|^2/2] \geqslant \pi(\boldsymbol{\theta}_0)$ and $(\pi(\boldsymbol{\theta}_0))^{-1} = \exp[O(p \log p)]$ by Condition (P). Since $\log n$ and $\log p$ are of the same order and $\log \det \mathbf{F}$ is at most of the order $p \log p$ [vide Condition (R)], the result follows by choosing c sufficiently large.

Proof of Lemma 2.5. By using the fact that $\|\mathbf{\Delta}_n\| = O_p(\sqrt{p})$, for large c and n, the left hand side of (2.18) can be bounded by $\Pr\{Y > cp/2\}$, with probability close to one, where Y has a central chi-square distribution with p degrees of freedom. The rest is merely a consequence of the standard large deviation estimates associated with the chi-square distribution; see Bahadur [1], for example.

ACKNOWLEDGMENT

The author is grateful to Professor J. K. Ghosh for suggesting this problem and for many fruitful discussions. Suggestions of the referee led to an improvement in the presentation. Most of the work was done when the author was a post doctoral fellow in the Indian Statistical Institute, Calcutta, supported by a grant from the National Board of Higher Mathematics, India.

REFERENCES

1. R. R. Bahadur, "Some Limit Theorems in Statistics," SIAM, Pennsylvania, 1971.
2. R. H. Berk, Consistency and asymptotic normality of MLE's for exponential models, *Ann. Math. Statist.* **43** (1972), 193–204.
3. P. Bickel and J. Yahav, Some contributions to the asymptotic theory of Bayes solutions, *Z. Wahrsch. Verw. Gebiete* **11** (1969), 257–275.
4. F. Chritley, P. Marriott, and M. Salmon, Preferred point geometry and statistical manifolds, *Ann. Statist.* **21** (1993), 1197–1224.

5. P. Diaconis and D. Freedman, On the Bernstein–von Mises theorem with infinite dimensional parameters, Technical Report 492, University of California, Berkeley, 1997.
6. M. Gasparini, Bayes Nonparametrics for biased sampling and density estimation, Unpublished Ph.D. thesis, University of Michigan, 1992.
7. S. Ghosal, Normal approximation to the posterior distribution for generalized linear models with many covariates, *Math. Methods Statist*. **6** (1997), 332–348.
8. S. Ghosal, Asymptotic normality of posterior distributions in high-dimensional linear models, *Bernoulli* **5** (1999), 315–331.
9. S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi, Posterior consistency of Dirichlet mixtures in density estimation, *Ann. Statist*. **27** (1999), 143–158.
10. S. Ghosal, J. K. Ghosh, and T. Samanta, On convergence of posterior distributions, *Ann. Statist*. **23** (1995), 2145–2152.
11. S. Ghosal, J. K. Ghosh, and A. W. van der Vaart, Convergence rates of posterior distribution, *Ann. Statist*. **28** (2000) (to appear).
12. U. Grenander, "Abstract Inference," Wiley, New York, 1981.
13. S. J. Haberman, Maximum likelihood estimates in exponential response models, *Ann. Statist*. **5** (1977), 815–841.
14. P. Huber, Robust regression: asymptotics, conjectures, and Monte Carlo, *Ann. Statist*. **1** (1973), 799–821.
15. R. A. Johnson, Asymptotic expansions associated with posterior distribution, *Ann. Math. Statist*. **42** (1970), 1241–1253.
16. L. Le Cam, On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, *Univ. California Publ. in Stat*. **1** (1953), 277–330.
17. M. S. Pinsker, Optimal filtration of square integrable signals in Gaussian noise, *Problems Inform. Transmission* **16** (1980), 120–133.
18. S. Portnoy, Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I: Consistency, *Ann. Statist*. **12** (1984), 1298–1309.
19. S. Portnoy, Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II: Normal approximation, *Ann. Statist*. **13** (1985), 1403–1417.
20. S. Portnoy, Asymptotic behavior of the empiric distribution of M-estimated residuals from a regression models with many parameters, *Ann. Statist*. **14** (1986), 1152–1170.
21. S. Portnoy, Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity, *Ann. Statist*. **16** (1988), 356–366.
22. W. H. Wong and X. Shen, Probability inequalities for likelihood ratios and convergence rates of sieve MLEs, *Ann. Statist*. **23** (1995), 339–362.
23. L. H. Zhao, Bayesian aspects of some nonparametric problems, Technical Report, University of Pennsylvania, 1998.